Final AFOSR Project Performance Report

DeLiang Wang
(Principal Investigator)

The Ohio State University

February 2012

This PI was awarded the AFOSR grant "Sequential organization and room reverberation for speech segregation" (Grant No.: FA9550-08-1-0155). The project was funded for the period of 2/1/08 to 11/30/11 with the total amount of $874K. This report summarizes the progress made throughout the project period.

1. RESEARCH PROGRESS

The human auditory system has the remarkable ability to segregate and follow a speaker in a noisy background. The perceptual theory of auditory scene analysis (ASA) provides a comprehensive account on how this ability is achieved. Inspired by the auditory scene analysis account, computational studies of speech segregation have made substantial advances in recent years. Despite these advances, speech segregation remains an extremely challenging, unsolved computational problem. Among major challenges are sequential organization, i.e. how to group the speech sounds of the same speaker across time, and room reverberation, i.e. how to achieve robustness against distortions caused by surface reflections of sound. Guided by ASA principles, this project primarily aimed to address these two challenges.

Consistent with the stated objectives of the project, the project made substantial progress along the following four directions. First, we have developed a tandem algorithm that performs pitch tracking and voiced speech segregation iteratively. Second, we have proposed a multipitch tracking algorithm for noisy and reverberant speech, which is then used in a novel, supervised learning approach to segregation of voiced speech in reverberant environments. Third, we have produced a method for unvoiced speech segregation by first removing voiced speech and periodic components, and then grouping unvoiced speech segments by analyzing their spectral characteristics. Fourth, we have proposed two algorithms for sequential organization, an unsupervised clustering algorithm applicable to monaural recordings and a binaural algorithm that integrates monaural and binaural analyses. In addition, we have conducted speech intelligibility tests that firmly establish the effectiveness of binary masking for improving human speech understanding in noisy backgrounds.

The major findings along the above directions are described in more detail in the following five subsections.

1.1 Tandem Algorithm

Natural speech contains both voiced and unvoiced portions, and voiced portions account for about 75-80% of spoken English. Voiced speech is characterized by periodicity (or harmonicity), which has been used as a primary cue in many computational auditory scene analysis (CASA) systems for segregating voiced speech. Despite considerable advances in voiced speech separation, the performance of current CASA systems is still limited by pitch (F0) estimation errors and residual noise. Various methods for robust pitch estimation have been proposed; however, robust pitch estimation under low signal-to-noise (SNR) situations still poses a significant challenge. Since the difficulty of robust pitch estimation stems from noise interference, it is desirable to remove or attenuate interference before pitch estimation. On the other hand, noise removal depends on accurate pitch estimation. As a result, pitch estimation and voice separation become a "chicken and egg" problem.

We believe that a key to resolve the above dilemma is the observation that one does not need the entire target signal to estimate pitch (a few harmonics can be adequate), and without perfect pitch one can still segregate some target signal. Thus, we suggest a strategy that estimates target pitch and segregates the target in tandem. The idea is that we first obtain a rough estimate of target pitch, and then use this estimate to segregate the target speech. With the segregated target, we should generate a better pitch estimate and can use it for better segregation, and so on. In other words, we propose a new algorithm that achieves pitch estimation and speech segregation jointly and iteratively. We call this method a *tandem algorithm* because it alternates between pitch estimation and speech segregation. This idea was present in a rudimentary form in the 2004 system developed by Hu and Wang (published in *IEEE Transactions on Neural Networks*) for voiced speech segregation which contained two iterations. Besides this idea, novel methods are proposed for segregation and pitch estimation; in particular, a classification based approach is proposed for pitch-based grouping.

The separation part of the tandem system aims to identify the *ideal binary mask* (IBM). With a time-frequency (T-F) representation, the IBM is a binary matrix along time and frequency where 1 indicates that the target is stronger than interference in the corresponding T-F unit and 0 otherwise. To simplify notations, we refer to T-F units labeled 1 and those labeled 0 as *unmasked* and *masked* units, respectively. We have suggested that the IBM is a primary goal for CASA, and it has since been used as a measure of ceiling performance in many speech separation studies. Psychoacoustic studies provide strong evidence that the IBM leads to large improvements of human speech intelligibility in noise (see Sect. 1.5).

Our tandem algorithm first generates an initial estimate of pitch contours and binary masks for up to two sources; a pitch contour refers to a consecutive set of pitches that is considered to be produced by the same sound source. In the initialization step, we first generate up to two estimated pitch periods in each time frame. Since T-F units dominated by a periodic signal tend to have high cross-channel correlations of filter responses or response envelopes, we only consider T-F units with high cross-channel correlations in this estimation. The algorithm then improves the estimation of pitch contours and masks

in an iterative manner. In the iterative estimation, we first re-estimate each pitch contour from its associated binary mask. A key step in this estimation is to expand estimated pitch contours based on temporal continuity, i.e., using reliable pitch points to estimate potential pitch points at neighboring frames. Then we re-estimate the mask for each pitch contour as follows. First, we compute the probability of each T-F unit dominated by the corresponding source of a pitch contour k. Then we estimate the mask for contour k according to the obtained probabilities.

Systematic evaluation shows that the tandem algorithm extracts a majority of target speech without including much interference, and it performs substantially better than previous systems for either pitch extraction or voiced speech segregation. As an example, Figure 1 shows the output from the tandem algorithm in response to a two-talker mixture, which consists of a set of pitch contours and their associated simultaneous streams. A paper describing the tandem algorithm was published in a 2010 paper by G. Hu and D.L. Wang, entitled "A tandem algorithm for pitch estimation and voiced speech segregation," in *IEEE Transactions on Audio, Speech and Language Processing.*

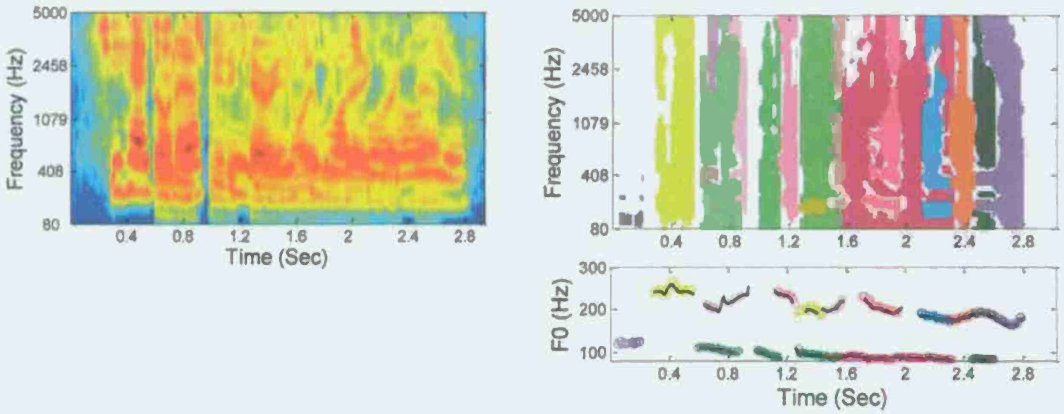

Figure 1. Tandem algorithm for pitch tracking and pitch-based grouping. **Left**: Cochleagram of the 0 dB mixture of a male and a female utterance. **Right**: Simultaneous streams (above) and their associated pitch contours (below), where different colors indicate different streams and their corresponding pitch contours. Ground-truth pitch contours are shown as curves and detected pitch points as circles.

1.2 Pitch Tracking and Pitch-based Grouping in Reverberant Environments

A robust pitch detection algorithm (PDA) is needed for many applications including CASA, prosody analysis, speech enhancement/separation, speech recognition, and speaker identification. Designing such an algorithm is challenging due to harmonic distortions brought about by acoustic interference and room reverberation. Numerous PDAs have been developed to detect a single pitch track under clean or modestly noisy conditions. The presumption of a single pitch track, however, puts limitations on the background noise in which PDAs perform. A multipitch tracker is required when the interfering sound also contains harmonic structure (e.g., background music or another

voice). A number of studies have investigated detecting multiple pitches simultaneously, including the tandem algorithm described above.

Room reverberation smears the characteristics of pitch (i.e., harmonic structure) in speech and thus makes the task of pitch determination more difficult. The performance of existing systems degrades substantially in reverberant environments. Little research has attempted to design and evaluate a multipitch tracker for reverberant speech signals, and what constitutes true pitch is even unclear in these conditions.

We have developed a multipitch tracking algorithm for both noisy and reverberant environments. First, we suggest a method to extract ground truth pitch for reverberant speech and use it as the reference for performance evaluation. After front-end processing, reliable channels are chosen based on cross-channel correlation and they constitute the summary correlogram for mid-level pitch representation. A pitch salience function is defined from which the pitch score of the observed correlogram given a pitch state is derived. The IBM is employed to divide selected channels into mutually exclusive groups, each corresponding to an underlying harmonic source. A hidden Markov model (HMM) integrates these pitch scores and searches for the best pitch state sequence. Our algorithm can reliably detect single and double pitch contours in noisy and reverberant conditions. Quantitative evaluations show that our approach significantly outperforms existing ones, particularly in reverberant conditions. This work was published in a 2011 paper by Z. Jin and D.L. Wang, entitled "HMM-based multipitch tracking for noisy and reverberant speech," in *IEEE Transactions on Audio, Speech and Language Processing*.

Speech segregation in reverberant environments is a very challenging problem. A monaural (one-microphone) solution is highly desirable in many important applications, e.g. as a frontend for automatic speech recognition and hearing aid design in noisy backgrounds. Numerous methods have been developed for monaural speech enhancement. These methods assume stationary or quasi-stationary interference and thus have intrinsic limitations in dealing with a general acoustic background. Model based approaches have been proposed to perform monaural segregation. However, none of these methods have been tested in reverberant conditions. Indeed, few studies have addressed the monaural speech segregation problem in room reverberation. One such study applied inverse filtering to partially counteract the smearing effect of reverberation on harmonic structure before segregation. However, the inverse filter is very sensitive to room configuration.

We have proposed a segregation system for reverberant speech by employing a supervised learning approach to classify harmonic cues in order to achieve robust segregation performance against reverberant effects. This supervised classification approach is used in conjunction with the above multipitch tracking algorithm developed for reverberant mixtures. We find that the classification approach continues to yield good performance when pitch-based features are extracted from estimated pitch, indicating good generalization. We further propose a novel unit labeling strategy for the time frames in which an interference pitch is also detected. Specifically, we train a multilayer perceptron (MLP) for target speech and a second MLP to model a variety of periodic interference. Because the MLP output estimates the posterior probability of the modeled source, a labeling criterion that compares the probabilities of the two underlying sources is expected to perform more reliably than one based on the posterior probability of only the target source. Here, we devise a likelihood ratio test to select the correct MLP model

for the interference. Experimental results show that the proposed system performs robustly in different types of interference and various reverberant conditions, and has a significant advantage over existing systems.

To show the results of our system, we construct an evaluation corpus using room impulse responses (RIRs) recorded in real rooms. Wechoose two acoustic rooms with the reverberation time (T_{60}) equal to 0.3 and 0.5 s, respectively. In each room, two RIRs corresponding to two omnidirectional microphones are selected for generating reverberant mixtures. The evaluation corpus is constructed by mixing TIMIT utterances with 15 different types of interference, which are classified into three categories: 1) those with no pitch, 2) those with some pitch qualities, and 3) other speech utterances, such that the segregation performance can be evaluated differently in each category. Fig. 2 presents and compares the SNR gains of the proposed system, the tandem algorithm and a spectral subtraction method in real rooms. The proposed algorithm performs significantly better than the other two algorithms in all reverberant conditions. These results demonstrate that our trained classifiers generalize well to real environments.

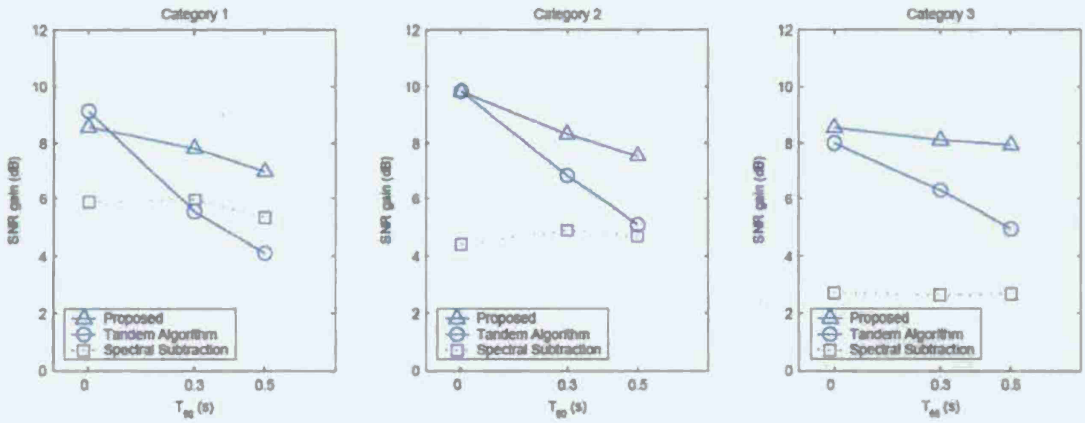

Figure 2. Comparison of SNR gain among the proposed method, the tandem algorithm and spectral subtraction in real rooms. The three panels (from left to right) indicate three different categories of interferences.

The reverberant speech separation system was recently published in another 2011 paper by Z. Jin and D.L. Wang, entitled "Reverberant speech segregation based on multipitch tracking and classification," in *IEEE Transactions on Audio, Speech and Language Processing*.

1.3. Unvoiced Speech Segregation

While a lot of efforts have been made in CASA to segregate voiced speech from monaural mixtures, unvoiced speech segregation remains a big challenge. Unvoiced speech lacks the pitch cue (or harmonic structure); in addition, due to its relatively weak energy unvoiced speech is highly susceptible to interference. As a subset of consonants, unvoiced speech consists of unvoiced fricatives, unvoiced stops and the unvoiced affricates. In our previous AFOSR project, we studied the unvoiced speech segregation

problem and successfully extracted a majority of unvoiced speech from nonspeech interference. Specifically, we utilized onset and offset cues to extract unvoiced speech segments. Acoustic-phonetic features are then used to separate unvoiced speech from nonspeech interference in a classification stage.

In a recent study, we have proposed a CASA-based system to estimate the unvoiced IBM, hence segregating unvoiced speech. We come up with the idea of spectral subtraction based segmentation and propose a conceptually and computationally simpler framework for segregation. First, our system segregates voiced speech by using the tandem algorithm. Considering that the task here is to segregate only unvoiced speech, we remove voiced speech as well as the periodic components in interference based on segregated voiced speech and cross-channel correlation. After the periodic part is removed, the mixture should only consist of unvoiced speech and aperiodic interference. Then unvoiced speech segregation occurs in two stages: segmentation and grouping. In segmentation, we first estimate the interference energy in unvoiced intervals by averaging the mixture energy in masked units in neighboring voiced intervals. Estimated noise energy is then used to perform spectral subtraction to generate unvoiced T-F segments.

In the grouping stage, unvoiced speech segments are extracted based on either simple thresholding or Bayesian classification. The simple thresholding method turns out to be quite effective, and works as follows. The energy of unvoiced speech often concentrates in the middle and high frequency ranges. This property, however, is not shared by nonspeech interference. To explore spectral characteristics of unvoiced speech and noise segments, we analyze their energy distributions with respect to frequency. Specifically, lower and upper frequency bounds of a segment are used to represent its frequency span. A statistical analysis reveals that unvoiced speech segments tend to reside at high frequencies while interference segments dominate at low frequencies. Interference is effectively removed at high frequencies during segmentation probably because the corresponding noise estimate is relatively accurate due to weak voiced speech at these frequencies. Based on this analysis and acoustic-phonetic characteristics of unvoiced speech, we can simply select segments with a lower bound higher than 2 kHz or an upper bound higher than 6 kHz as unvoiced speech and remove others as noise.

A systematic comparison shows the proposed system outperforms the system developed in the previous project over a wide range of input SNR levels. In addition, segmentation based on spectral subtraction is simpler and faster than multiscale onset-offset analysis, and grouping based on simple thresholding does not need supervised training. Our CASA based approach also performs substantially better than two representative speech enhancement methods, indicating the effectiveness of a grouping stage. Figure 3 shows the comparative results. In the case of using only spectral subtraction, the largest gap is about 10 dB when the input SNR is -5 dB and the gap is about 1.8 dB as the input SNR increases to 15 dB. The Wiener-as algorithm performs worse than spectral subtraction. The unvoiced speech separation system was recently published in a 2011 paper by K. Hu and D.L. Wang, entitled "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," in *IEEE Transactions on Audio, Speech and Language Processing.*

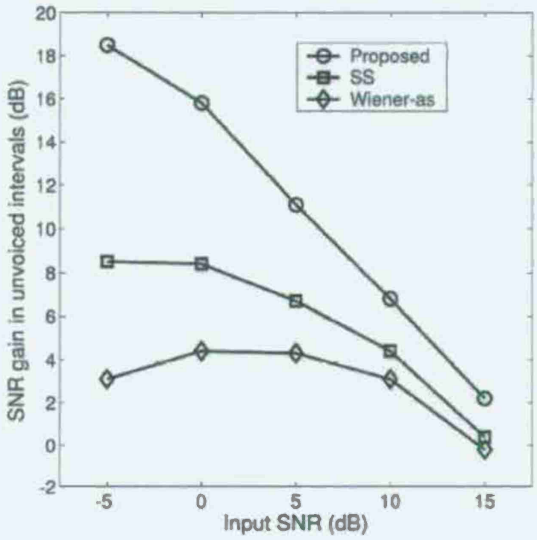

Figure 3. Comparison with two representative speech enhancement methods at different input SNR levels. The two methods are spectral subtraction (SS) and *a priori* SNR based Wiener algorithm (Wiener-as).

1.4. Sequential Organization

An unsupervised monaural approach

Cochannel speech refers to the mixture of two speech signals transmitted simultaneously in a single channel. Under cochannel conditions, two talkers are usually not aware of each other and the resultant speech mixture often has a large amount of overlap. Such a condition poses a real difficulty to speech separation and recognition. Previous studies on cochannel speech separation employ model-based methods. One approach extends the framework of single-speaker identification to the two-talker case and group speech components by maximizing the joint speaker recognition score. Similarly, HMMs have been employed to model speakers and speech is separated by coupling segregation and recognition. Model-based methods can achieve satisfactory performance when trained models match those of participating speakers. However, this condition is often not met in practice.

We aim to separate cochannel speech in an unsupervised way that requires no prior speaker knowledge. We first decompose the speech mixture into T-F segments, which are further grouped across frequency to form simultaneous streams. Each simultaneous stream is mainly dominated by a single speaker and continuous in time. Different simultaneous streams are generally separated in time and how to group them into individual speakers is the task of sequential grouping. Sequential grouping resembles speaker clustering but has two unique challenges. First, simultaneous streams consist of spectrally separated components while speech sections in speaker clustering contain spectrally whole frames. Second, a simultaneous stream is much shorter than a speech section in speaker clustering.

We have developed a clustering-based method for unsupervised sequential organization of speech. We introduce a cepstral feature generated directly from the mixture based on T-F masking. To deal with unvoiced speech, we first employ an onset/offset analysis to extract T-F segments from the whole speech mixture. Then, the portions of segments overlapping with segregated voiced speech are removed, and the remaining portions are re-segmented to produce unvoiced speech segments. For each unvoiced segment, we calculate its overlap with the complementary portions of the segregated voiced speech of each speaker. We then label each unvoiced segment by comparing overlap patterns. Without using any prior models, our method is completely unsupervised and applies to both voiced speech and unvoiced speech.

The key step in our approach is to formulate sequential organization of voiced speech as a problem of unsupervised clustering: voiced simultaneous streams are clustered into two speaker groups. Our clustering objective function is based on the ratio of between- and within-group distances:

$$O(\mathbf{g}) = tr(\frac{\mathbf{S}_B(\mathbf{g})}{\mathbf{S}_W(\mathbf{g})})$$

where $\mathbf{g}$ is a hypothesized binary label vector for all voiced simultaneous streams, and $\mathbf{S}_W(\mathbf{g})$ and $\mathbf{S}_B(\mathbf{g})$ are within- and between-group scatter matrices, respectively. Specifically, according to $\mathbf{g}$, simultaneous streams are divided into two groups and we pool the masked cepstral features in individual groups to calculate $\mathbf{S}_W(\mathbf{g})$ and $\mathbf{S}_B(\mathbf{g})$. The trace operator in the above equation is used to measure the group distance, which can be interpreted as the ratio of the between- and within-group scatter matrices along the eigenvector dimensions. Sequential grouping is then achieved by maximizing $O(\mathbf{g})$.

Evaluations and comparisons show that our unsupervised method outperforms a model-based method in terms of speech segregation. A short report describing our unsupervised approach was published by K. Hu and D.L. Wang, entitled "An approach to sequential grouping in cochannel speech," in the *Proceedings of* 2011 *ICASSP*. A comprehensive version is currently under review for journal publication.

A binaural approach in room reverberation

Most existing approaches to binaural or sensor-array based speech segregation have relied exclusively on localization cues embedded in the differences between signals recorded by multiple microphones. These approaches may be characterized as spatial filtering (or beamforming), which enhances the signal from a specific direction. Spatial filtering approaches can be very effective in certain acoustic conditions. On the other hand, beamforming has well known limitations. Chief among them is substantial performance degradation in reverberant environments.

Our study proposes an alternative framework that integrates monaural and binaural analysis to achieve robust localization and segregation of voiced speech in reverberant environments. In the language of CASA, our proposed system uses monaural cues to achieve simultaneous organization. This allows locally extracted, unreliable binaural cues

to be integrated over large T-F regions. Integration over such regions enhances localization robustness in reverberant conditions and in turn, we use robust localization to achieve *sequential organization*.

Our computational framework is partly motivated by psychoacoustic studies suggesting that binaural cues may not play a dominant role in simultaneous organization, but are important for sequential organization. Further, human listeners are able to effectively localize multiple sound sources in reverberant environments, and recent analysis suggests that localization may be facilitated by monaural grouping, rather than localization acting as a fundamental grouping cue in ASA.

The proposed system integrates monaural and binaural analysis to achieve segregation of voiced speech. In the first stage of the system, the tandem algorithm is used to form *simultaneous streams* from the T-F units of the *better ear* signal. By better ear signal, we mean the signal in which the input SNR is higher, as determined from the signals before mixing. A simultaneous stream is thought to be dominated by the same source.

Binaural cues are extracted that measure differences in timing and level between corresponding T-F units of the left and right ear signals. A set of trained, azimuth-dependent likelihood functions are then used to map from timing and level differences to cues related to source location. Azimuth cues are integrated within simultaneous streams in a probabilistic framework to achieve sequential organization and to estimate the underlying source locations. The output of the system is a set of streams, one for each source in the mixture, and the azimuth angles of the underlying sources.

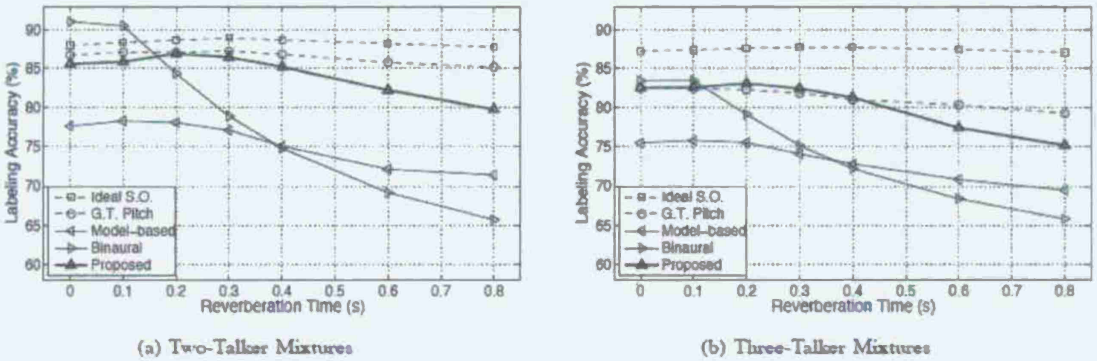

Figure 4. Labeling accuracy of the proposed and comparison systems shown as a function of reverberation time for (a) two-talker and (b) three-talker mixtures. Ceiling performance measures are shown with dashed lines. In the figure, S.O. stands for sequential organization, and G.T. stands for ground truth.

In Figure 4, we show the performance of the proposed system, a model-based comparison system, the binaural system and the ideal sequential organization scheme on the two- and three-talker mixtures. The performance achieved by ideal sequential organization indicates the quality of the monaural simultaneous organization by the tandem algorithm. Any decrease below 100% reflects that the simultaneous streams are not exclusively dominated by target or interference. Besides a model-based system, another main comparison is made to a binaural only system (see Fig. 4). It is clear from

Figure 4 that the proposed system represents a significant improvement over the binaural system, and that the margin between the two increases as a function of reverberation time. This sequential organization approach was published in a 2010 paper by J. Woodruff and D.L. Wang, entitled "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," in *IEEE Transactions on Audio, Speech and Language Processing.*

1.5 Speech Intelligibility Tests

What should a sound segregation system aim for? With a T-F representation, we have suggested the use of the ideal binary time-frequency mask as a primary goal of CASA. The idea behind the IBM is to retain the T-F regions of a mixture where the target is relatively strong, and discard the remaining regions. Specifically, an ideal mask is a binary matrix whose value is one for a T-F unit where the local SNR within the unit exceeds a threshold or local criterion (LC), and zero otherwise. The notion of the ideal binary mask is directly motivated by the auditory masking phenomenon that, roughly speaking, a stronger signal within a critical band masks a weaker signal.

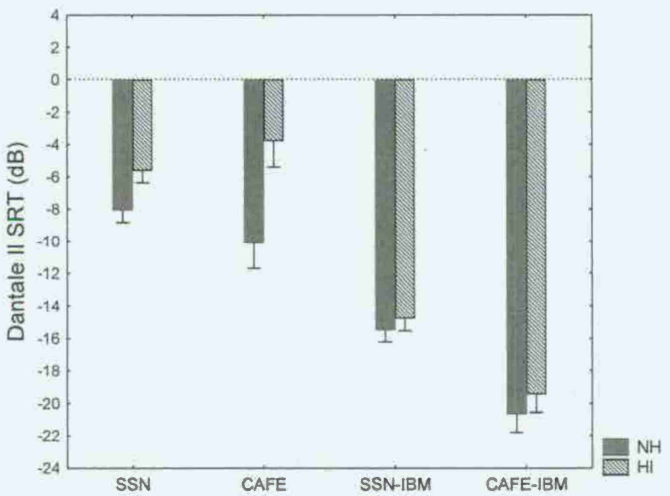

Figure 5. Speech reception thresholds before and after ideal masking for NH and HI listeners with SSN and cafeteria noises. Error bars indicate 95% confidence intervals of the means.

A key question for the ideal binary mask as the goal of speech segregation is whether the IBM leads to speech intelligibility improvements. We have conducted a study to evaluate the speech intelligibility improvements of ideal masking for both normal-hearing (NH) and hearing-impaired (HI) listeners. Our study fixed LC to -6 dB and measured SRT (speech reception threshold) effects using both a speech-shaped noise (SSN) and a cafeteria noise. The speech material consisted of Danish sentences from the Dantale II corpus. Figure 5 shows the results where "SSN" and "CAFE" indicate the unprocessed conditions and "SSN-IBM" and "CAFE-IBM" the ideal binary masking conditions. As shown in the figure, we have observed a 7.4-dB SRT reduction (i.e. improvement) for NH

listeners and a 9.2-dB reduction for HI listeners with the SSN background, and a 10.5-dB reduction for NH listeners and a 15.6-dB reduction for HI listeners with the cafeteria background. The observed SRT improvements for the cafeteria noise are significantly larger than for SSN, suggesting that ideal masking is more effective for modulated noise than for stationary noise. Strikingly, ideal masking makes the intelligibility performances for HI listeners and NH listeners comparable. This study was published in a 2009 paper by Wang et al., entitled: "Speech intelligibility in background noise with ideal binary time-frequency masking," in the *Journal of the Acoustical Society of America.*

In a related study, we have made a surprising finding that listeners achieve nearly perfect speech recognition from pure noise that is turned on and off according to the IBM. This process of turning on or off noise is illustrated in Figure 6. IBM-gated noise produces almost perfectly intelligible speech. This result is surprising as the information encoded in binary gains is greatly reduced from original speech. The result was published in a 2008 paper by Wang et al., entitled: "Speech perception of noise with binary gains," in the *Journal of the Acoustical Society of America.*

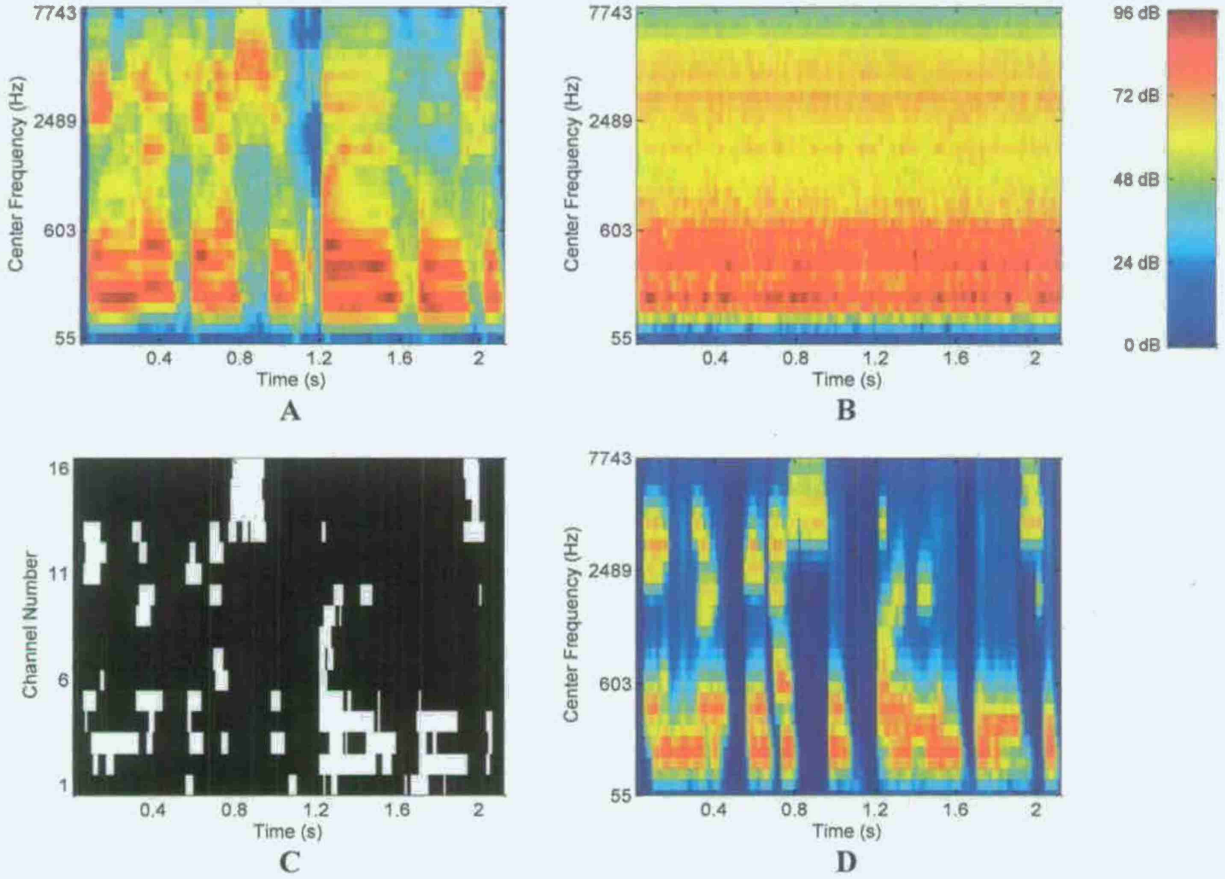

Figure 6. **A** and **B** show the cochleagrams of a sentence and a noise, respectively. **C** shows the IBM with 16 frequency channels, where a white pixel indicates 1 and a black indicates 0. **D** shows the noise in **B** gated by the IBM in **C**. Note the similarity between **A** and **D**.

These studies provide compelling evidence that ideal binary masking produces large intelligibility gains. The intelligibility benefit is even larger for HI listeners than for NH listeners, and for modulated noise than for steady noise. As far as CASA is concerned, the clear implication is that the solution is not about estimating the target signal, as has been attempted in the field for many years, but about classifying the mixture. The shift of problem formulation from estimation to classification has major ramifications because, now, the speech segregation (cocktail party) problem is open to a plethora of powerful machine learning techniques. Subsequent developments have finally led to classification algorithms that can improve human speech intelligibility in noise, a goal that has eluded the community for decades.

2. OTHER INFORMATION

2.1 Development of Human Resources

The project in various stages has supported four doctoral students as graduate research assistants: Yipeng Yi, Zhaozhang Jin, John Woodruff, and Ke Hu. The support enabled Li and Jin to complete their doctoral studies, and Woodruff and Hu to nearly finish theirs.

Li's work is mostly on separation of musical sounds where pitch plays a major role. As part of his doctoral study, he established the optimality of the ideal binary mask. Li gave a formal treatment that, under certain conditions, the IBM is indeed the optimal binary mask in terms of SNR among all the binary masks. His research also shows that IBMs are close in performance to ideal ratio masks which are closely related to the Wiener filter, the theoretically optimal linear filter. Li's dissertation was completed in 2008. An executive summary of the dissertation is given in Appendix 1. His dissertation is available online at:

http://etd.ohiolink.edu/view.cgi?acc_num=osu1211994188.

Jin's work deals with on segregation of voiced speech in reverberant environments, and two pieces of his work are described in Sect. 1.2. Jin's research led to a Ph.D. dissertation entitled "Monaural speech segregation in reverberant environments" completed in 2010. An executive summary of the dissertation is given in Appendix 2. His dissertation is available online at:

http://etd.ohiolink.edu/view.cgi/Jin%20Zhaozhang.pdf?osu1279141797.

Woodruff's work integrates monaural grouping and location-based grouping using binaural cues. Sect. 1.4 describes one piece of his work on using localization for sequential organization in reverberant environments. The theme of combining monaural and binaural analyses has resulted in state-of-the-art algorithms for sound localization and speech segregation in room reverberation.

Hu's doctoral study has produced an unvoiced speech separation algorithm introduced in Sect. 1.3, and an unsupervised approach to sequential organization introduced in Sect. 1.4. Currently, he is completing a study on elevating the performance of model-based speech segregation by leveraging the progress on unsupervised grouping.

This grant has helped the PI to update a graduate-level course entitled "Computational

audition", and enhance the existing graduate-level courses "Introduction to Neural Networks" and "Brain Theory and Neural Networks". Additionally, the PI has participated in a great deal of curriculum and seminar activity for training undergraduate students.

2.2 Awards/Honors

The PI received the 2008 Helmholtz Award from the International Neural Network Society. This award is given to an individual annually for outstanding achievement in sensation and perception. The PI was selected to be an IEEE Distinguished Lecturer for the three years of 2010-2012. In addition, the PI received the 2010 Lumley Research Award from the OSU College of Engineering. This is the fourth consecutive time the PI received this recognition.

John Woodruff received the 2011 Best Poster Award at the Department of Computer Science and Engineering graduate research exhibition. In addition, he just received a 2012 Graduate Student Research Award from the department.

2.3 Transition or Collaborative Activities

The PI continues collaborating with researchers at AFRL in Dayton (Drs. Douglas Brungart and Brian Simpson). The collaboration has led to a study examining the effects of speaker characteristics and the number of talkers in ideal binary masking. A paper summarizing the results was published by the *Journal of the Acoustical Society of America* in 2009.

The PI had a 2-year contract from AFRL/IF in Rome (through RADC) to study speaker recognition in noisy conditions, which was completed in June 2011. For the project, we applied the results from this AFOSR project to perform speech segregation in order to achieve robust speaker recognition. The research supported by the AFOSR project uncovered a new auditory feature, called Gammatone Frequency Cepstral Coefficient (GFCC). We demonstrated that GFCC is more robust to background noise than widely used MFCC (Mel-Frequency Cepstral Coefficient). We transmitted the program code for GFCC extraction as well as a state-of-the-art robust speaker identification system to RADC/AFRL. The contact at RADC is Dr. Brett Smolensky.

Our results on speaker separation and identification have also been transferred to SPI, a small business company located in Rockville, Maryland. We provided to them the program code to assist in their project on audio clustering and speaker recognition. The SPI contact is Dr. Chiman Kwan.

Kuzer, a small-business company located in Seattle, Washington, has partnered with the PI in winning a Phase I/II STTR project funded by AFOSR. The Phase II just started, and the project aims to develop a prototype speech separation system that can improve speech intelligibility of human listeners in background noise. The project represents a direct effort applying the algorithms developed under AFOSR support to solve an important task with many important applications.

Our results of integrating monaural grouping in sound localization in reverberant environments (see Sect. 1.4) were transitioned to Oticon A/S, which is located in Copenhagen, Denmark, and is a major hearing aid manufacturer in the world. Through a 6-month visit in 2010 by John Woodruff, Oticon designed a system that utilizes monaural segregation to enhance the directionality of input signals received by a hearing aid. The Oticon contact is Dr. Ulrik Kjems.

2.4 Publications

Journal articles

Hu G. and Wang D.L. (2008): "Segregation of unvoiced speech from nonspeech interference," *Journal of the Acoustical Society of America*, vol. 124, pp. 1306-1319.

Wang D.L., Kjems U., Pedersen M.S., Boldt J.B., and Lunner T. (2008): "Speech perception of noise with binary gains," *Journal of the Acoustical Society of America*, vol. 124, pp. 2303-2307.

Srinivasan S. and Wang D.L. (2008): "A model for multitalker speech perception," *Journal of the Acoustical Society of America*, vol. 124, pp. 3213-3224.

Wang D.L. (2008): "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, pp. 332-353.

Li Y. and Wang D.L. (2009): "On the optimality of ideal binary time-frequency masks," *Speech Communication*, vol. 51, pp. 230-239.

Jin Z. and Wang D.L. (2009): "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 625-638.

Wang D.L., Kjems U., Pedersen M.S., Boldt J.B., and Lunner T. (2009): "Speech intelligibility in background noise with ideal binary time-frequency masking," *Journal of the Acoustical Society of America*, vol. 125, pp. 2336-2347.

Li Y. and Wang D.L. (2009): "Musical sound separation based on binary time-frequency masking," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, Article ID 130567, 10 pages.

Li Y., Woodruff J., and Wang D.L. (2009): "Monaural musical sound separation based on pitch and common amplitude modulation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1361-1371.

Brungart D.S., Chang P.S., Simpson B.D., and Wang D.L. (2009): "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and speaker number" *Journal of the Acoustical Society of America*, vol. 125, pp. 4006-4022.

Shao Y. and Wang D.L. (2009): "Sequential organization of speech in computational auditory scene analysis," *Speech Communication*, vol. 51, pp. 657-667.

Kjems U., Boldt J.B., Pedersen M.S., Lunner T., and Wang D.L. (2009): "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," *Journal of the Acoustical Society of America*, vol. 126, pp. 1415-1426.

Shao Y., Srinivasan S., Jin Z., and Wang D.L. (2010): "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech and Language*, vol. 24, pp. 77-93.

Srinivasan S. and Wang D.L. (2010): "Robust speech recognition by integrating speech separation and hypothesis testing," *Speech Communication,* vol. 52, pp. 72-81.

Woodruff J. and Wang D.L. (2010): "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1856-1866.

Narayanan A. and Wang D.L. (2010): "Robust speech recognition from binary masks," *Journal of the Acoustical Society of America Express Letters*, vol. 128, pp. EL217-222.

Hu G. and Wang D.L. (2010): "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 2067-2079.

Jan T., Wang W., Wang D.L. (2011): "A multistage approach to blind separation of convolutive speech mixtures," *Speech Communication*, vol. 53, pp. 524-539.

Jin Z. and Wang D.L. (2011): "HMM-based multipitch tracking for noisy and reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1091-1102.

Hu K. and Wang D.L. (2011): "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1600-1609.

Jin Z. and Wang D.L. (2011): "Reverberant speech segregation based on multipitch tracking and classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 2328-2337.

Hsu C.-L., Wang D.L., and Jang J.-S.R., Hu K. (2012): "A tandem algorithm for singing pitch extraction and voice separation from music accompaniment," *IEEE Transactions on Audio, Speech, and Language Processing*, in press.

Woodruff J. and Wang D.L. (2012): "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Transactions on Audio, Speech, and Language Processing*, in press.

Zhao X., Shao Y. and Wang D.L. (2012): "CASA-based robust speaker identification," *IEEE Transactions on Audio, Speech, & Language Processing*, in press.

Book chapters

Wang D.L. and Hu G. (2008): "Cocktail party processing," In: Zurada J.M., Yen G.G., and Wang J. (ed.), Computational Intelligence: Research Frontiers, Springer, Berlin, pp. 333-348.

Roman N. and Wang D.L. (2008): "Binaural speech segregation," In: Hänsler E. and Schmidt G. (ed.), Speech and Audio Processing in Adverse Environments, Springer, Berlin, pp. 525-549.

Narayanan A. and Wang D.L. (2012): "Computational auditory scene analysis and automatic speech recognition," In: Virtanen T., Raj B. and Singh R. (ed.), *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, Chichester U.K., in press.

Conference papers

Li Y. and Wang D.L. (2008): "Musical sound separation using pitch-based labeling and binary time-frequency masking," *Proceedings of ICASSP*-08, pp. 173-176.

Li Y. and Wang D.L. (2008): "On the optimality of ideal binary time-frequency masks," *Proceedings of ICASSP*-08, pp. 3501-3504.

Shao Y. and Wang D.L. (2008): "Robust speaker identification using auditory features and computational auditory scene analysis," *Proceedings of ICASSP*-08, pp. 1589-1592.

Boldt J.B., Kjems U., Pedersen M.S., Lunner T., and Wang D.L. (2008): "Estimation of the ideal binary mask using directional systems," *Proceedings of International Workshop on Acoustic Echo and Noise Control* (IWAENC-08), 4 pages.

Woodruff J., Li Y., and Wang D.L. (2008): "Resolving overlapping harmonics for monaural musical sound separation using pitch and common amplitude modulation," *Proceedings of ISMIR-08*, pp. 538-543.

Jan T., Wang W., Wang D.L. (2009): "A multistage approach for blind separation of convolutive speech mixtures," *Proceedings of ICASSP-09*, pp. 1713-1716.

Woodruff J. and Wang D.L. (2009): "On the role of localization cues in binaural segregation of reverberant speech," *Proceedings of ICASSP-09*, pp. 2205-2208.

Hu K. and Wang D.L. (2009): "Incorporating spectral subtraction and noise type for unvoiced speech segregation," *Proceedings of ICASSP-09*, pp. 4425-4428.

Shao Y., Jin Z., Wang D.L., and Srinivasan S. (2009): "An auditory-based feature for robust speech recognition," *Proceedings of ICASSP-09*, pp. 4625-4628.

Jin Z. and Wang D.L. (2009): "Learning to maximize signal-to-noise ratio for reverberant speech segregation," *Proceedings of ICASSP-09*, pp. 4689-4692.

Jin Z. and Wang D.L. (2010): "A multipitch tracking algorithm for noisy and reverberant speech," *Proceedings of ICASSP-10*, pp. 4218-4221.

Woodruff J. and Wang D.L. (2010): "Integrating monaural and binaural analysis for localization of multiple sound sources in noisy and reverberant conditions," *Proceedings of ICASSP-10*, pp. 2706-2709.

Kjems U., Boldt J.B., Pedersen M.S., Lunner T., and Wang D.L. (2010): "Speech intelligibility of ideal binary masked mixtures," *Proceedings of EUSIPCO-10*, pp. 1909-1913.

Hu K. and Wang D.L. (2010): "Unsupervised sequential organization for cochannel speech separation," *Proceedings of INTERSPEECH-10*, pp. 2790-2793.

Hu K. and Wang D.L. (2010): "Unvoiced speech segregation based on CASA and spectral subtraction," *Proceedings of INTERSPEECH-10*, pp. 2786-2789.

Woodruff J., Prabhavalkar R., Fosler-Lussier E., and Wang D.L. (2010): "Combining monaural and binaural evidence for reverberant speech segregation," *Proceedings of INTERSPEECH-10*, pp. 406-409.

Woodruff J. and Wang D.L. (2011): "Directionality-based speech enhancement for hearing aids," *Proceedings of ICASSP-11*, pp. 297-300.

Hsu C.-L., Wang D.L. and Jang J.R. (2011): "A trend estimation algorithm for singing pitch detection in musical recordings," *Proceedings of ICASSP-11*, pp. 393-396.

Narayanan A. and Wang D.L. (2011): "On the use of ideal binary masks for improving phonetic classification," *Proceedings of ICASSP-11*, pp. 4632-4635.

Hu K. and Wang D.L. (2011): "An approach to sequential grouping in cochannel speech," *Proceedings of ICASSP-11*, 4636-4639.

Narayanan A., X. Zhao, Wang D.L., and Fosler-Lussier E. (2011): "Robust speech recognition using multiple prior models for speech reconstruction," *Proceedings of ICASSP-11*, pp. 4800-4803.

Han K. and Wang D.L. (2011): "An SVM based classification approach to speech separation," *Proceedings of ICASSP-11*, pp. 5212-5215.

Zhao X., Shao Y., and Wang D.L. (2011): "Robust speaker identification using a CASA front-end," *Proceedings of ICASSP-11*, pp. 5468-5471.

Woodruff J. and Wang D.L. (2012): "Binaural speech segregation based on pitch and azimuth tracking," *Proceedings of ICASSP-12*, in press.

Han K. and Wang D.L. (2012): "On generalization of classification based speech separation," *Proceedings of ICASSP-12*, in press.

Hu K. and Wang D.L. (2012): "SVM-based separation of unvoiced-voiced portions in cochannel speech," *Proceedings of ICASSP-12*, in press.

Appendix 1. Executive Summary of Yipeng Li's Ph.D. Dissertation

Monaural musical sound separation attempts to isolate one or more sound sources from a single-channel polyphonic signal. The main motivation of this study is the capability of the human auditory system in organizing an acoustic mixture into different perceptual streams which correspond to different sound sources. The underlying perceptual process is called auditory scene analysis (ASA) and it has inspired the development of computational auditory scene analysis (CASA).

A recent development in CASA is the establishment of ideal binary masks (IBM) as a major goal for CASA. The IBM has several desirable properties as an objective of CASA systems and one of them is the purported optimality with respect to signal-to-noise ratio (SNR) among all the binary masks. However, this optimality has not been rigorously addressed. This dissertation gives a formal treatment on this issue and clarifies the conditions for the IBM to be optimal. This dissertation also shows that IBMs are close in performance to ideal ratio masks which are closely related to the Wiener filter, the theoretically optimal linear filter. As a result, the IBM is adopted as our computational goal for musical sound separation when binary masking is considered.

Pitch is a primary cue in the perceptual organization of sounds. Since the majority of musical sounds are pitched, this dissertation is centered on pitch-based sound separation. The first system aims to separate singing voice from music accompaniment and features an effective algorithm for detecting the pitch contours of singing voice in the presence of other musical sounds. The system consists of three stages. The singing voice detection stage partitions and classifies an input into vocal and non-vocal portions. For vocal portions, the predominant pitch detection stage detects the pitch contours of the singing voice and then the separation stage uses the detected pitch contours to group the time-frequency segments of the singing voice. Quantitative results show that the system performs the separation task successfully.

The second system attempts to separate instrument sounds from a polyphonic signal. This system focuses on addressing the problem of overlapping harmonics, a major difficulty in musical sound separation. To make reliable binary decisions on which source has stronger energy in an overlapping region, the contextual information of sounds is utilized based on the assumption that sounds from the same source tend to have similar spectral envelopes. Quantitative results show that this strategy can help binary decisions in overlapping regions and consequently improve the SNR performance of separation.

To achieve higher separation quality, a sinusoidal modeling-based separation system is developed with the emphasis on resolving overlapping harmonics. This system also utilizes contextual information of sounds: harmonics of the same source have correlated amplitude envelopes. This is known as common amplitude modulation in ASA. Another observation is that the phase change of harmonics can be predicted from pitch points. These two observations are incorporated in a least-squares estimation framework for separation. An effective technique is introduced to improve the accuracy of pitch estimation and make the system applicable to practical applications. Quantitative evaluation of the proposed system shows that it performs significantly better than existing monaural musical sound separation systems.

Appendix 2. Executive Summary of Zhaozhang's Ph.D. Dissertation

Room reverberation is a major source of signal degradation in real environments. While listeners excel in "hearing out" a target source from sound mixtures in noisy and reverberant conditions, simulating this perceptual ability remains a fundamental challenge. The goal of this dissertation is to build a computational auditory scene analysis (CASA) system that separates target voiced speech from its acoustic background in reverberant environments. A supervised learning approach to pitch-based grouping of reverberant speech is proposed, followed by a robust multipitch tracking algorithm based on a hidden Markov model (HMM) framework. Finally, a monaural CASA system for reverberant speech segregation is designed by combining the supervised learning approach and the multipitch tracker.

Monaural speech segregation in reverberant environments is a particularly challenging problem. Although inverse filtering has been proposed to partially restore the harmonicity of reverberant speech before segregation, this approach is sensitive to specific source/receiver and room configurations. Assuming that the true target pitch is known, the first study in this dissertation leads to a novel supervised learning approach to monaural segregation of reverberant voiced speech, which learns to map a set of pitch-based auditory features to a grouping cue encoding the posterior probability of a time-frequency (T-F) unit being target dominant given observed features. The study devises a novel objective function for the learning process, which directly relates to the goal of maximizing signal-to-noise ratio. The model trained using this objective function yields significantly better T-F unit labeling. A segmentation and grouping framework is utilized to form reliable segments under reverberant conditions and organize them into streams. Systematic evaluations show that the proposed approach produces very promising results under various reverberant conditions and generalizes well to new utterances and new speakers.

Multipitch tracking in real environments is critical for speech signal processing. Determining pitch in both reverberant and noisy conditions is another difficult task. In the second study, a robust algorithm is proposed for multipitch tracking in the presence of background noise and room reverberation. A new channel selection method is utilized to extract periodicity features. The study derives pitch scores for each pitch state, which estimate the likelihoods of the observed periodicity features given pitch candidates. An HMM integrates these pitch scores and searches for the best pitch state sequence. This algorithm can reliably detect single and double pitch contours in noisy and reverberant conditions.

Building on the first two studies, the final study proposes a CASA approach to monaural segregation of reverberant voiced speech, which performs multipitch tracking of reverberant mixtures and supervised classification. Speech and nonspeech models are separately trained, and each learns to map pitch-based features to the posterior probability of a T-F unit being dominated by the source with the given pitch estimate. Because interference can be either speech or nonspeech, a likelihood ratio test is introduced to select the correct model for labeling corresponding T-F units. Experimental results show that the proposed system performs robustly in different types of interference and various reverberant conditions, and has a significant advantage over existing systems.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
28-02-2012	Final performance report	2/2008 - 11/2011

4. TITLE AND SUBTITLE
Sequential organization and room reverberation for speech segregation

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-08-1-0155

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
DeLiang Wang (Principal Investigator)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Ohio State University
Research Foundation
1960 Kenny Road
Columbus, OH 43210-1063

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. Willard Larkin
AFOSR/NL, Room 713
4015 Wilson Blvd.
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-OSR-VA-TR-2012-0068

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is limited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Inspired by the perceptual account of auditory scene analysis, significant advances were made in speech segregation in recent years. Despite these advances, two major challenges remained: sequential organization and room reverberation. This project aimed to address these two challenges. Substantial progress has been made along the following directions. First, a tandem algorithm was developed that performs pitch tracking and voiced speech segregation iteratively. Second, a multipitch tracking algorithm was proposed for noisy and reverberant speech, which was then used in a novel, supervised learning approach to segregation of voiced speech in reverberant environments. Third, a method was suggested for unvoiced speech segregation by first removing voiced speech and periodic components, and then grouping unvoiced speech segments through analyzing their spectral characteristics. Two algorithms were proposed for sequential organization, an unsupervised clustering algorithm applicable to monaural recordings and a binaural algorithm that integrates monaural and binaural analyses. In addition, speech intelligibility tests were conducted and their results firmly establish the effectiveness of binary masking for improving human speech recognition in noisy backgrounds.

15. SUBJECT TERMS
Computational auditory scene analysis, speech segregation, sequential organization, room reverberation, computational audition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			DeLiang Wang
Unclassified	Unclassified	Unclassified	Unclassified	21	19b. TELEPHONE NUMBER *(Include area code)* 614-292-6827